Learning causal network structure from multiple (in)dependence models

Tom Claassen
Radboud University, Nijmegen
tomc@cs.ru.nl

Tom Heskes
Radboud University, Nijmegen
tomh@cs.ru.nl

Abstract

We tackle the problem of how to use information from multiple (in)dependence models, representing results from different experiments, including background knowledge, in causal discovery. We introduce the framework of a causal system in an external context to derive a connection between strict conditional independencies and causal relations between variables. Constraint-based causal discovery is shown to be decomposable into a candidate pair identification and a subsequent elimination step that can be applied separately from different models. The result is the first principled, provably sound method that is able to infer valid causal relations from different experiments in the large sample limit. We present a possible implementation that shows what results can be achieved and how it might be extended to other application areas.

1 Introduction

Discovering causal relations from observational data is an important, ubiquitous problem in science. In many application areas there is data available from many different but related experiments. Results obtained from one data set are often used to either corroborate or challenge results from another. Yet how to reconcile apparently contradictory information from multiple sources, including background knowledge, into a single, more informative model remains a long-standing open problem.

Constraint-based methods like the FCI-algorithm (Spirtes et al., 2000) are provably correct in the large sample limit, even in the presence of latent variables; the same holds for Bayesian methods like the greedy search algorithm GES (Chickering, 2002) (with additional post-processing steps to handle hidden confounders). Both are defined in terms of modeling a single data set and have no principled means to relate to results from other sources in the process. Recent developments, like the ION-algorithm by Tillman et al. (2008), show that it is possible to integrate multiple, partially overlapping data sets, provided they originate from *identical* experiments. These are still essentially single model learners as they assume there is one underlying structure that can account for *all* observed dependencies in the different models. In practice there are often inconsistencies between data sets, precisely because the experimental circumstances were not identical. The way out is to distinguish between causal dependencies internal to the system under investigation and merely contextual dependencies.

In section 4 we show that causal discovery can be decomposed into two separate steps: a conditional independency to identify a pair of possible causal relations (one of which is true), and then a conditional dependency to eliminate one of the candidates, leaving the other. The two steps are independent and rely only on the observed (in)dependencies between a subset of variables. As a result conclusions remain valid, even when taken from different models.

2 Graphical model preliminaries

First a few familiar notions from graphical model theory used throughout the article.

A *directed graph* $\mathcal{G}$ is a pair $\langle \mathbf{V}, \mathbf{E} \rangle$, where $\mathbf{V}$ is a set of vertices or nodes and $\mathbf{E}$ is a set of edges

between pairs of nodes. Edges are represented by arrows $X \to Y$, where node X is the *parent* of Y and Y is a *child* of X. Two vertices are *adjacent* in $\mathcal{G}$ if there is an edge between them. A *path* $\pi = \langle V_0, \ldots, V_n \rangle$ between V_0 and V_n in $\mathcal{G}$ is a sequence of distinct vertices such that for $0 \leq i \leq n-1$, V_i and V_{i+1} are adjacent in $\mathcal{G}$. A *directed path* is a path that is traversed entirely in the direction of the arrows. A *directed acyclic graph* (DAG) is a directed graph that does not contain a directed path from any node to itself. A vertex X is an *ancestor* of Y (and Y is a *descendant* of X) if there is a directed path from X to Y in $\mathcal{G}$ or if $X = Y$. A vertex Z is a *collider* on a path $\pi = \langle \ldots, X, Z, Y, \ldots \rangle$ if it contains the subpath $X \to Z \leftarrow Y$, otherwise it is a *noncollider*. A *trek* is a path that does not contain any collider.

For disjoint sets of vertices $\mathbf{X}$, $\mathbf{Y}$ and $\mathbf{Z}$ in a DAG $\mathcal{G}$, $\mathbf{X}$ is *d-connected* to $\mathbf{Y}$ conditional on $\mathbf{Z}$ (possibly empty), iff there exists an unblocked path $\pi = \langle X, \ldots, Y \rangle$ between some $X \in \mathbf{X}$ and some $Y \in \mathbf{Y}$, i.e. such that every collider on π is an ancestor of some $Z \in \mathbf{Z}$ and every noncollider on π is not in $\mathbf{Z}$. If not, then all such paths are blocked, and $\mathbf{X}$ is said to be *d-separated* from $\mathbf{Y}$ given $\mathbf{Z}$. Note that in a DAG $\mathcal{G}$, an unblocked path π between two vertices X and Y cannot be blocked by conditioning on a node Z that is not on the path, and that a blocked path can only be unblocked by conditioning on (descendants of) all colliders on the path; see (Pearl, 2000; Spirtes et al., 2000) for more details.

Let p be a probability distribution over a set of variables $\mathbf{V}$, and let $\mathbf{X}$, $\mathbf{Y}$ and $\mathbf{Z}$ denote three disjoint subsets of $\mathbf{V}$, then an *(in)dependence model* is a set of (in)dependence statements that hold in p of the form '$\mathbf{X}$ is independent of $\mathbf{Y}$ given $\mathbf{Z}$', denoted $\mathbf{X} \perp\!\!\!\perp \mathbf{Y} \mid \mathbf{Z}$, and/or '$\mathbf{X}$ is dependent of $\mathbf{Y}$ given $\mathbf{Z}$', denoted $\mathbf{X} \not\perp\!\!\!\perp \mathbf{Y} \mid \mathbf{Z}$, with set $\mathbf{Z}$ possible empty. (In)dependence models are often compactly and intuitively represented in the form of a graphical model (directed, undirected or other), in combination with a criterion to link the structure of the graph to the implied (in)dependencies, similar to the d-separation for DAGs. We will pose no restrictions on shape or type of the (in)dependence models considered in this article, other than that they are internally consistent.

3 Modeling the system

This section introduces the framework of a causal system in an external context to model experiments, as well as a number of assumptions adopted throughout the rest of this article.

3.1 Causal DAG

A causal DAG is a graphical model in the form of a DAG where the arrows represent direct causal interactions between variables in a system. A prime characteristic of a causal structure is the so-called *Manipulation Principle* (Spirtes et al., 2000), which boils down to the fact that changing/manipulating a variable will affect all and only its descendants in the causal DAG. In this article we will not concern ourselves with the interpretation of causality any further; for that the reader is referred to (Cartwright, 2004; Williamson, 2005). Instead, we simply assume that the systems we consider can be represented by some underlying causal DAG over a great many observed and unobserved nodes. In a causal DAG $\mathcal{G}_C$ there is a *causal relation* from variable X to variable Y iff there is a directed path π from X to Y in $\mathcal{G}_C$, otherwise it is a *noncausal relation*. A direct link $X \Rightarrow Y$ in the graph $\mathcal{G}_C$ means that there is a causal path from X to Y that is not mediated by any other node in $\mathcal{G}_C$.

The ubiquitous **causal Markov condition** links the structure of a causal graph to its probabilistic concomitant, (Pearl, 2000): two variables X and Y in a causal DAG $\mathcal{G}_C$ are dependent given a set of nodes $\mathbf{Z}$, iff they are connected by a path π in $\mathcal{G}_C$ that is unblocked given $\mathbf{Z}$. An immediate consequence is that there is a dependence $X \not\perp\!\!\!\perp Y$ iff there is a trek between X and Y in the causal DAG.

Another common assumption which we will adopt throughout the article is the **causal faithfulness condition** which implies that all and only the conditional independence relations entailed by the causal Markov condition applied to the true causal DAG will hold in the joint probability distribution over the variables in $\mathcal{G}_C$.

For an in-depth discussion of the justification of and connection between these assumptions in causal inference, see (Pearl, 2000; Spirtes et al., 2000; Zhang and Spirtes, 2008).

3.2 Experimental context

Random variation in a system (a.k.a. 'error terms' in a structural equation model (SEM)), corresponds to the impact of unknown external variables (Pearl, 2000). Some external factors may be actively controlled, as for example in clinical trials, or passively observed as the natural embedding of a system in its environment. We refer to both observational and controlled studies as *experiments*. If there are external factors that affect two or more variables in a system simultaneously, then this can lead to an observed dependency that is not part of the system (a.k.a. 'correlated errors' in SEMs). Both can be represented by modeling this external environment explicitly as a set of unknown, hypothetical context nodes that causally affect the system under scrutiny. We introduce:

Definition 1. The *external context* of a causal DAG $\mathcal{G}_C$, denoted $\mathcal{G}_E$, is an additional set of mutually independent nodes $\mathbf{U}$ in combination with links from every $U \in \mathbf{U}$ to one or more nodes in $\mathcal{G}_C$.

The total causal structure of an experiment on a causal system $\mathcal{G}_C$ in external context $\mathcal{G}_E$ is then denoted by $\mathcal{G}_T = \{\mathcal{G}_E + \mathcal{G}_C\}$. The context only introduces arrows from nodes in $\mathcal{G}_E$ to $\mathcal{G}_C$ which can never result in a cycle if there was not one in $\mathcal{G}_C$ already (there are no links between nodes in $\mathcal{G}_E$). Therefore, the structure of an experiment $\mathcal{G}_T$ is also a causal DAG. In this paradigm different experiments become variations in context of an *invariant* causal system.

Figure 1 depicts a causal system in two different contexts (double lined arrows indicate direct causal relations; dashed circles represent unobserved variables). The experiment on the right hand side will result in an observed dependency between variables A and B, whereas the one on the left will not.

Here we only focus on the (in)dependence relations $\mathcal{I}(\mathbf{V} \subset \mathcal{G}_C)$ that exist in the joint probability distribution $P(\mathbf{V})$ over the observed subset of variables for a given causal experiment $\{\mathcal{G}_E + \mathcal{G}_C\}$. With this we can state the goal of causal discovery from multiple models as: "Given experiments with unknown total causal structures $\mathcal{G}_T = \{\mathcal{G}_E + \mathcal{G}_C\}$, $\mathcal{G}'_T = \{\mathcal{G}'_E + \mathcal{G}_C\}$, etc., and corresponding (in)dependence models $\mathcal{I}(\mathbf{V} \subset \mathcal{G}_C)$, $\mathcal{I}'(\mathbf{V}' \subset \mathcal{G}_C)$, etc., which variables are connected by a directed path in $\mathcal{G}_C$?". We assume that in each experiment the large sample limit distributions are known and have been used to obtain categorical statements about probabilistic (in)dependencies between sets of nodes. As stated, we will also always assume that the causal Markov and causal faithfulness condition are satisfied.

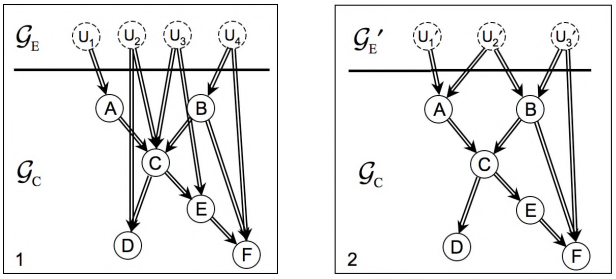

Figure 1: A causal system $\mathcal{G}_C$ in different experiments

4 Conditional (in)dependence in causal systems

Given the problem statement above, we need a way to combine (in)dependence statements from different models in order to identify causal relations in the underlying causal structure $\mathcal{G}_C$ that is assumed to be at the heart of all of them. Methods like FCI and GES tackle this reconstruction problem in terms of properties that are optimal or minimal w.r.t. a model for a given experiment, but this gives no means to relate results from different models. Another approach, taken in the ION algorithm, is to use ancestral graph theory (Richardson and Spirtes, 2002) to establish what probabilistic (in)dependencies will be observed in a causal experiment for different subsets of observed variables, and then use this to find relations that must be shared by all. But this still does not allow to combine

results from *different* experiments, like in fig. 1.

A way out of this predicament comes courtesy of a remarkable fact that so far (to the best of our knowledge) has escaped detection in causal research: there is a fundamental connection between causality and a certain type of conditional independence, that applies regardless of the encompassing model. This connection will enable us to bring together results from arbitrary experiments in a method for causal discovery from multiple (in)dependence models (section 6). To exclude irrelevant independencies we first introduce the following notion:

Definition 2. Two nodes X and Y are *strictly conditionally (in)dependent* given a set of nodes $\mathbf{Z}$, iff X is conditionally (in)dependent of Y given a *minimal* set of nodes $\mathbf{Z}$.

We denote a strict (in)dependence statement by placing it in square brackets. The *minimal* in the definition implies that the relation does not hold for any proper subset of the (possibly empty) set $\mathbf{Z}$, e.g. a strict conditional independence $[X \perp\!\!\!\perp Y \,|\, \mathbf{Z}]$ implies both $X \perp\!\!\!\perp Y \,|\, \mathbf{Z}$ and $\forall \mathbf{Z}' \subsetneq \mathbf{Z} : X \not\perp\!\!\!\perp Y \,|\, \mathbf{Z}'$. It aims to capture the notion that it is really the entire set $\mathbf{Z}$ that makes X and Y independent. The relevance of this notion lies in the fact that, in a causal system, certain causal relations between three variables X, Y and Z can *never* result in an observed strict conditional independence $[X \perp\!\!\!\perp Y \,|\, Z]$, no matter what the context is.

Example 1. For the causal system $\mathcal{G}_C$ in fig.2a (two variables $X \Rightarrow Z$ with no causal links to or from a variable Y), there is no context $\mathcal{G}_E$ that can result in $[X \perp\!\!\!\perp Y \,|\, Z]$: if there are no directed paths from Z to X and Y then $X \not\perp\!\!\!\perp Y$ implies that X and Y are d-connected by directed paths $\langle U, \ldots, X \rangle$ and $\langle U, \ldots, Y \rangle$ that do not contain Z. But then conditioning on Z cannot block these paths, ergo not $X \perp\!\!\!\perp Y \,|\, Z$. This does not apply to causal system in fig.2b: for the indicated context $\mathcal{G}_E$ the strict conditional independence relation $[X \perp\!\!\!\perp Y \,|\, Z]$ will be observed.

A quick survey shows that all causal structures over three nodes that *can* lead to an observed $[X \perp\!\!\!\perp Y \,|\, Z]$ have a direct causal link from Z to X and/or Y.

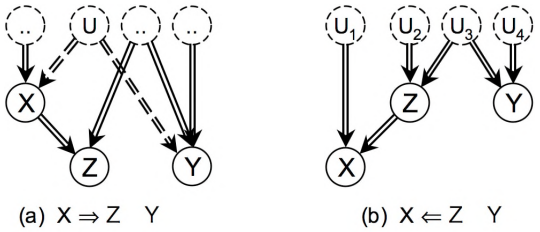

Figure 2: Causal systems $\mathcal{G}_C$ that: (a) cannot, and (b) depending on the context $\mathcal{G}_E$ *can* lead to an observed strict conditional independence relation $[X \perp\!\!\!\perp Y \,|\, Z]$.

We can generalize this result to sets of nodes:

Theorem 1. *In an experiment with causal structure $\mathcal{G}_T = \{\mathcal{G}_E + \mathcal{G}_C\}$, a strict conditional independence $[X \perp\!\!\!\perp Y \,|\, \mathbf{Z}]$ implies causal links $Z \Rightarrow X$ and/or $Z \Rightarrow Y$ from every $Z \in \mathbf{Z}$ to X and/or Y in $\mathcal{G}_C$.*

Proof. We construct a directed path for an arbitrary $Z_1 \in \mathbf{Z}$ to either X or Y. Z_1 must be a noncollider on some path π_1 connecting X and Y given all the other nodes $\mathbf{Z}_{\setminus Z_1}$. Follow π_1 in the direction of the arrows (choose either branch if Z_1 has two outgoing arrows along π_1) until either X or Y or a collider that is an ancestor of one of the remaining nodes in $\mathbf{Z}_{\setminus Z_1}$ is encountered. If X or Y is found first then a directed path has been found and we are done. If not then we can go on from the collider along π_1 to its descendant node $Z_2 \in \mathbf{Z}_{\setminus Z_1}$. This node in turn must be a noncollider on some other path π_2 that d-connects X and Y given all nodes $\mathbf{Z}_{\setminus Z_2}$. Again this path can be followed in the direction of the arrows until either X or Y or a collider that is ancestor of one of the nodes in $\mathbf{Z}_{\setminus \{Z_1, Z_2\}}$ is encountered. (This cannot be one of the previous nodes since that would imply the existence of a directed path.) We can continue, and as long as neither X nor Y is reached we will find new nodes from $\mathbf{Z}$ until all have been encountered. At that point the final node will lie on a trek connecting X and Y that can no longer be blocked by any other node in $\mathbf{Z}$, and therefore will have a directed path to X or Y. By construction that means there is also a directed path from Z_1 to either X or Y in $\mathcal{G}_C$, which implies a causal relation $Z_1 \Rightarrow X$ and/or $Z_1 \Rightarrow Y$. □

This theorem recognizes conditional independence as the 'local signature' of causality. It is not difficult to see that for a single Z the causal link to X or Y is also *unconfounded* (no hidden common parent). This plays an important role in calculating the magnitude of causal effects, e.g. via the front-door criterion (Pearl, 2000).

A similar result exists for conditional dependence and noncausal relations, something we already knew for v-structures (unshielded colliders $X \rightarrow Z \leftarrow Y$) from (Spirtes et al., 2000), although not in the general form given here:

Theorem 2. *Let X, Y, $\mathbf{Z}$ and $\mathbf{W}$ be disjoint (sets of) nodes in an experiment with causal structure $\mathcal{G}_T = \{\mathcal{G}_E + \mathcal{G}_C\}$. If there is a conditional independence $X \perp\!\!\!\perp Y \,|\, \mathbf{W}$ and a minimal set $\mathbf{Z}$ such that $X \not\perp\!\!\!\perp Y \,|\, \{\mathbf{W} \cup \mathbf{Z}\}$, then there are no causal links $Z \Rightarrow X$, $Z \Rightarrow Y$, and/or $Z \Rightarrow W$ from any $Z \in \mathbf{Z}$ to any X, Y and/or $W \in \mathbf{W}$ in $\mathcal{G}_C$.*

Proof. We show it holds for arbitrary $Z \in \mathbf{Z}$. In short: Z must be a (descendant of a) collider on a path connecting X and Y (otherwise it would not be needed to unblock the path); any directed path from Z to a W implies that conditioning on Z is not needed when already conditioning on W. No directed paths from Z to $\mathbf{W}$ implies that if there existed a directed path from Z to X or Y then it cannot be blocked by any W; neither can it be blocked by any $\mathbf{Z}_{\setminus Z}$ (otherwise $\mathbf{Z}$ is not minimal). But then such a path would make Z a noncollider on an unblocked path between X and Y given $\mathbf{Z}_{\setminus Z}$, contradicting minimality. □

With the addition of $\mathbf{W}$ the theorem also applies to unshielded colliders where X and Y are not independent. We need one more result that is particularly useful to eliminate direct links between variables in a causal model:

Theorem 3. *In an experiment with causal structure $\mathcal{G}_T = \{\mathcal{G}_E + \mathcal{G}_C\}$, every conditional independence $X \perp\!\!\!\perp Y \,|\, \mathbf{Z}$ implies the absence of causal paths $X \Rightarrow Y$ or $X \Leftarrow Y$ in $\mathcal{G}_C$ between X and Y that are not mediated by nodes in $\mathbf{Z}$.*

Proof. If there did exist causal paths between X and Y not mediated by $\mathbf{Z}$ then conditioning on $\mathbf{Z}$ would not block all directed paths (let alone treks) between X and Y, so then $X \not\perp\!\!\!\perp Y \,|\, \mathbf{Z}$. □

5 Identifying causal relations

Theorem 1 and 2 together show that causal discovery can be decomposed into two *separate* steps: having a means of identifying a pair of links that harbors a causal relation as well as a means of eliminating a causal relation as the origin of an observed link, the obvious consequence is that this allows the positive identification of a definite causal relation.

Corollary 1. *In an experiment $\mathcal{G}_T = \{\mathcal{G}_E + \mathcal{G}_C\}$, if there exists a strict conditional independence $[X \perp\!\!\!\perp Y \,|\, \mathbf{Z}]$, then if there also exists a conditional independence $X \perp\!\!\!\perp V \,|\, \mathbf{W}$ and $\mathbf{Z}$ is a minimal set such that $X \not\perp\!\!\!\perp V \,|\, \{\mathbf{W} \cup \mathbf{Z}\}$, then there are causal links $Z \Rightarrow Y$ from every $Z \in \mathbf{Z}$ to Y in $\mathcal{G}_C$.*

Proof. By theorem 1 $[X \perp\!\!\!\perp Y \,|\, \mathbf{Z}]$ implies causal links from every $Z \in \mathbf{Z}$ to X and/or Y. The second condition, $X \perp\!\!\!\perp V \,|\, \mathbf{W}$ with $\mathbf{Z}$ minimal such that $X \not\perp\!\!\!\perp V \,|\, \{\mathbf{W} \cup \mathbf{Z}\}$, applies to theorem 2 and implies that there are no causal links from any $Z \in \mathbf{Z}$ to X. With all links from $\mathbf{Z}$ to X eliminated, the only remaining option is causal links $Z \Rightarrow Y$ from every $Z \in \mathbf{Z}$ to Y. □

To illustrate how these rules can be applied to infer causal links directly from observed (in)dependence relations, we look at two independence models (represented in figure 3 as CPAGs, see Appendix A), that are known, e.g. from the FCI-algorithm (Spirtes et al., 2000), to contain a definite causal link, and show how this also follows as a straightforward application of theorems 1 and 2.

Example 2. The aptly named *Y-structure* in the l.h.s. of fig. 3 plays an important role in causal discovery: every such substructure in a minimal independence model derived by the FCI-algorithm allows the identification of causal link $Z \Rightarrow Y$, i.e. a directed path from Z to Y is present in *all* possible causal DAGs corresponding to the observed distribution over the variables. Mani et al. (2006) investigated marginal Y-structures *embedded* in data sets. It

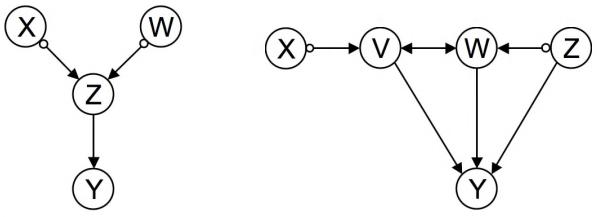

Figure 3: Two independence models in the form of a CPAG: the 'Y-structure' (left) and a discriminating path (right), both with a detectable causal link $Z \Rightarrow Y$ (arrow in CPAG); see examples for detailed description.

was shown that for any DAG, in the large sample limit, a consistent Bayesian scoring function (Heckerman et al., 1999) will assign a higher score to a structure with a direct link $Z \to Y$, when marginalizing over the variables, than to any structure without. These results are easily understood in terms of our theorems: any (embedded) Y-structure satisfies the relations $[X \perp\!\!\!\perp Y \mid Z]$ and $[X \not\perp\!\!\!\perp W \mid Z]$. By theorem 1, the first implies $Z \Rightarrow X$ or $Z \Rightarrow Y$, the second eliminates $Z \not\Rightarrow X$ by theorem 2, leaving $Z \Rightarrow Y$.

As another example, we look at the following important, but somewhat awkward, construct in causal inference: in a graph $\mathcal{G}$, a path $\pi = \langle X, \ldots, W, Z, Y \rangle$ is a *discriminating path* for Z if X is not adjacent to Y and all nodes between X and Z are colliders on π and parents of Y.

Example 3. The path $\pi = \langle X, V, W, Z, Y \rangle$ in the r.h.s. of figure 3 is a discriminating path for Z. The relevance of such a path for causal discovery lies in the fact that if $Z \to Y$ is present in the graph $\mathcal{G}$, then it is present in all members of the equivalence class of $\mathcal{G}$, and hence it corresponds to a definite causal link (Spirtes et al., 2000; Zhang, 2008). The causal implication of this discriminating path can also be understood in terms of the previous rules: by definition of π it follows that X and Y are strict conditionally independent given some set $\mathbf{Z}$ (otherwise they would be adjacent in $\mathcal{G}$). If there is a link $Z \to Y$, then Z (and all other nodes between X and Z on π) is necessarily part of any $\mathbf{Z}$ that will d-separate X and Y. Therefore, figure 3 implies $[X \perp\!\!\!\perp Y \mid Z \cup \{V, W\}]$ and $X \not\perp\!\!\!\perp Z \mid \varnothing$, which by theorems 1 and 3 implies $Z \Rightarrow Y$.

6 Causal relations from multiple models

As all three theorems (rules) in section 4 hold separately for experiments $\mathcal{G}_T$ *irrespective* of the context $\mathcal{G}_E$, it means that (non)causal results obtained in one experiment should also apply to another, provided the causal system $\mathcal{G}_C$ remains invariant. In that case, an algorithm implementing these rules should be able to construct a single, overall model of the causal relations that is more informative than any of the (in)dependence models separately.

For that we note that all noncausal information ('X does not cause Y') from rules (2) and (3) derives from single models in isolation, and so can be processed first and collected in a matrix of (non)causal relations found. Subsequent causal relations identified via rule (1) also imply reverse noncausal information, which in turn can lead to new causal relations. This suggests a repeated loop until no new information can be found. As input for the algorithm we use CPAGs (see Appendix A) as a concise and intuitive graphical representation of all invariant (in)dependence features in an observed distribution, e.g. as learned by the extended FCI-algorithm (Zhang, 2008). To convey all uncovered information about the underlying causal structure $\mathcal{G}_C$ we choose a *causal PAG* $\mathcal{G}$ as the output model: similar in form and interpretation to a CPAG, where a missing edge between variables corresponds to the absence of a direct causal path, every detected direct noncausal link $X \not\Rightarrow Y$ has an arrowhead at X in $\mathcal{G}$, every detected direct causal link $X \Rightarrow Z$ has a tail mark at X in $\mathcal{G}$, and circle marks represent unknown, possibly causal relations. A straightforward implementation is provided in algorithm 1.

To illustrate the algorithm, consider the CPAG models corresponding to two experiments on the l.h.s. of figure 4. Despite the different, even apparently contradictory, observed (in)dependence relations, the combined causal model on the r.h.s. is readily derived.
Starting from the fully connected graph, in the first loop over the models, rule (3) in line

Input : set of CPAGs $\mathcal{P}_i \in \mathbf{P}$
Output : causal graph $\mathcal{G}$

1: $\mathcal{G} \leftarrow$ fully connected graph with circle marks
2: $\mathbf{M}_C \leftarrow \mathbf{0}$ ▷ *empty set of (non-)causal relations*
3: **for all** $\mathcal{P}_i \in \mathbf{P}$ **do**
4: **for all** $(X, Y, Z) \in \mathcal{P}_i$, with no edge $X - Y$ **do**
5: $\mathbf{M}_C \leftarrow X \not\leftarrow Y, Y \not\leftarrow X$ if $X \perp\!\!\!\perp Y \,|\, \varnothing$ ▷ *Rule (3)*
6: **for all** $\mathbf{W} \in \{\mathcal{P}_i \setminus X, Y, Z\}$ **do**
7: **if** $X \perp\!\!\!\perp Y \,|\, \mathbf{W}$ **then**
8: $\mathcal{G} \leftarrow$ eliminate edge $X \multimap\!\!\!\!\circ\; Y$ ▷ *Rule (3)*
9: **if** $X \not\perp\!\!\!\perp Y \,|\, \{\mathbf{W} \cup Z\}$ **then**
10: $\mathbf{M}_C \leftarrow Z \not\leftarrow \{X, Y, \mathbf{W}\}$ ▷ *Rule (2)*
11: **end if**
12: **end if**
13: **end for**
14: **end for**
15: **end for**
16: $\mathcal{G} \leftarrow$ noncausal info in $\mathbf{M}_C$ ▷ *circles to arrowheads*
17: **repeat**
18: **for all** $\mathcal{P}_i \in \mathbf{P}$ **do**
19: **for all** $(X, Y) \in \mathcal{P}_i$, with no edge $X - Y$ **do**
20: **for all** $\mathbf{Z} \in \{\mathcal{P}_i \setminus X, Y\}$ **do**
21: **if** $[X \perp\!\!\!\perp Y \,|\, \mathbf{Z}]$ **and** $X \not\leftarrow Z \in \mathbf{M}_C$ **then**
22: $\mathbf{M}_C \leftarrow Z \Rightarrow Y$ **and** $Y \not\leftarrow Z$ ▷ *Rule (1)*
23: **end if**
24: **end for**
25: **end for**
26: **end for**
27: $\mathcal{G} \leftarrow$ (non)causal info in $\mathbf{M}_C$ ▷ *tails/arrowheads*
28: **until** no more new noncausal information found

Algorithm 1: Causal structure inference algorithm

8 eliminates all links except $A - C$, $B - C$, $B - F$, $C - D$, $C - E$ and $E - F$ (missing edges in input model). In the same loop, model 1 has $[A \not\perp\!\!\!\perp B \,|\, \{C/D/E/F\}]$ which by rule (3) in line 5 implies $A \not\leftarrow B$ and $B \not\leftarrow A$, and from which rule (2) in line 10 derives noncausal links $\{C/D/E/F\} \not\leftarrow \{A, B\}$ (for empty $\mathbf{W}$ in theorem 2) . In the subsequent repeated loop, lines 17-28, model 1 has $[A \perp\!\!\!\perp F \,|\, \{B, C\}]$ which by rule (1) in line 22 with the earlier $B \not\leftarrow A$, implies $B \Rightarrow F$. Similarly, $[C \perp\!\!\!\perp F \,|\, \{B, E\}]$ allows the conclusion $E \Rightarrow F$. Next, model 2 has $[A \perp\!\!\!\perp D \,|\, C]$ which, together with $C \not\leftarrow A$ implies $C \Rightarrow D$. Finally, from $[A \perp\!\!\!\perp E \,|\, C]$ follows $C \Rightarrow E$. After that the algorithm terminates at line 28 with the causal CPAG on the r.h.s. as the final output. (Figure 1 shows two contexts that can account for the observed dependencies in figure 4).

To the best of our knowledge, this is the first algorithm ever to perform such a derivation. The input in the form of CPAGs is convenient, but not essential: any (in)dependence

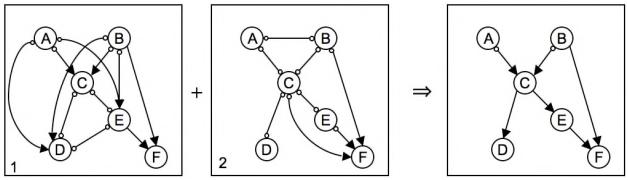

Figure 4: Independence models (in CPAG form) for two experiments, one resulting causal model (cf. fig.1).

model can be used with only minor alterations to the implementation. We could even directly incorporate (non)causal information from background knowledge in the first loop. In the current form the example derivation is almost instantaneous, but soon becomes unfeasible for larger networks. Also the set of observed variables can differ between input models, but with little overlap causal information may be lost if it cannot be transferred to the output graph when other information has eliminated that particular direct link. Nevertheless, all identified (non)causal relations remain valid. These problems can be addressed and significant improvements can be made, but that requires additional results and explication and will be postponed to another article.

7 Discussion

We have shown the first principled method to use information from *different* (in)dependence models in causal discovery. It is based on the discovery of a fundamental property that identifies (strict) conditional independence as the local signature of causality. All (non)causal relations uncovered this way are sound, provided the input models are valid. The number and individual size and origin of the input models are irrelevant and could include different experiments, specific background knowledge or hypothetical information. An exciting possibility is to use this approach in combination with recent developments that employ other properties of the distribution, e.g. non-Gaussianity (Shimizu et al., 2006) or nonlinear features (Hoyer et al., 2009), to detect causal relations.

The proposed algorithm is sound and works well on small models ($\lesssim 10$ nodes) with a rea-

sonable degree of overlap. In order to apply the method to larger, more realistic models with less overlap, further research should concentrate on the computational complexity of the search for (new) strict conditional independencies and ways to handle indirect causal information. If the input models become less reliable, for example when derived from real data sets where the large sample limit no longer applies, incorrect or inconsistent causal conclusions may occur. In that case, results might be generalized to quantities like 'the probability of a causal relation' based on the strength and reliability of the required conditional (in)dependencies in the available data.

Acknowledgments

This research was supported by VICI grant 639.023.604 from the Netherlands Organization for Scientific Research (NWO).

Appendix A. CPAGs

For a causal DAG the distribution over a subset of observed variables may not be faithfully representable by a DAG. A *complete partial ancestral graph* (CPAG) $\mathcal{P}$ represents the Markov equivalence class $[\mathcal{G}]$ of a DAG $\mathcal{G}$ when latent variables may be present (Zhang, 2008). It is a graph with either a tail '$-$' (signifying ancestorship), arrowhead '$\blacktriangleright$' (signifying non-ancestorship) or circle mark '$\circ$' at each end of an edge. There is a tail or arrowhead on an edge in $\mathcal{P}$ iff it is invariant in $[\mathcal{G}]$, otherwise it has a circle mark. Bi-directed edges $\blacktriangleleft\!\!\blacktriangleright$ in a CPAG indicate the presence of a latent common cause; arcs $\longrightarrow$ indicate a causal relation. The CPAG is unique and maximally informative for $[\mathcal{G}]$. An intuitive property of CPAGs is that two nodes X and Y are not connected by an edge iff there is some set $\mathbf{Z}$ such that $X \perp\!\!\!\perp Y \mid \mathbf{Z}$; see (Richardson and Spirtes, 2002; Zhang, 2008) for more information on how to read (in)dependencies directly from a CPAG using the m-separation criterion.

References

N. Cartwright. 2004. Causation: one word, many things. *Philosophy of Science*, (71):805–819.

D. Chickering. 2002. Optimal structure identification with greedy search. *Journal of Machine Learning Research*, 3(3):507–554.

D. Heckerman, C. Meek, and G. Cooper. 1999. A Bayesian approach to causal discovery. In *Computation, Causation, and Discovery*, pages 141–166.

P. Hoyer, D. Janzing, J. Mooij, J. Peters, and B. Schölkopf. 2009. Nonlinear causal discovery with additive noise models. In *Advances in Neural Information Processing Systems 21 (NIPS*2008)*, pages 689–696.

S. Mani, G. Cooper, and P. Spirtes. 2006. A theoretical study of Y structures for causal discovery. In *Proceedings of the 22nd Conference in Uncertainty in Artificial Intelligence*, pages 314–323.

J. Pearl. 2000. *Causality: models, reasoning and inference*. Cambridge University Press.

T. Richardson and P. Spirtes. 2002. Ancestral graph Markov models. *Ann. Stat.*, 30(4):962–1030.

S. Shimizu, P. Hoyer, A. Hyvärinen, and A. Kerminen. 2006. A linear non-Gaussian acyclic model for causal discovery. *Journal of Machine Learning Research*, 7:2003–2030.

P. Spirtes, C. Glymour, and R. Scheines. 2000. *Causation, Prediction, and Search*. The MIT Press, Cambridge, Massachusetts, 2nd edition.

R. Tillman, D. Danks, and C. Glymour. 2008. Integrating locally learned causal structures with overlapping variables. In *Advances in Neural Information Processing Systems, 21*.

J. Williamson. 2005. *Bayesian nets and causality: philosophical and computational foundations*. Oxford University Press, Oxford.

J. Zhang and P. Spirtes. 2008. Detection of unfaithfulness and robust causal inference. *Minds and Machines*, 2(18):239–271.

J. Zhang. 2008. On the completeness of orientation rules for causal discovery in the presence of latent confounders and selection bias. *Artificial Intelligence*, 172(16-17):1873 – 1896.